AIAA Guidance, Navigation, and Control Conf., Austin, TX, Aug., 2003.

A CLUSTER ANALYSIS TO CLASSIFY DAYS IN THE NATIONAL AIRSPACE SYSTEM

Bob Hoffman, Ph.D.,
Jimmy Krozel, Ph.D., Steve Penny
Metron Aviation, Inc.
131 Eldon St., Suite 200
Herndon, VA 20170

Anindya Roy, Ph. D.
Dept. of Mathematics and Statistics
Univ. of Maryland at Baltimore County
Baltimore, MD 21250

Karlin Roth, Ph.D.
NASA Ames Research Center
MS 210-10
Moffett Field, CA 94035

Abstract

Simulations spanning the entire National Airspace System (NAS) are of growing interest. For NAS simulations, researchers must select an appropriate set of days to validate their simulation models. However, within vast quantities of historical NAS data, certain days have had abnormal events that create unusual or anomalous traffic flow patterns – September 11, 2001 being the most notable. For such atypical days, researchers may wish to avoid them or reserve them for special analysis or modeling. Furthermore, the researcher is confronted with the problem of how to consolidate vast quantities of descriptive NAS data into a more easily interpretable description of daily operational behavior that can help in selecting appropriate days for simulation validation. In this paper, we design a NAS feature vector to characterize the NAS behavior for comparing across days. Cluster analysis is used to condense an initial collection of 65 aggregate, daily NAS variables into a more manageable set of variables. Each feature vector then represents NAS performance on its respective day. In a second but different cluster analysis, we rank NAS feature vectors (and therefore days) by levels of normality. Finally, we provide recommendations on how to use these "typical" days to validate NAS simulations.

Introduction

What defines a "typical" day in the National Airspace System (NAS)? More generally, is there a systematic way of categorizing days by the degree to which they are normal or abnormal? A closely related problem is to determine how to consolidate vast quantities of descriptive NAS data into a more easily interpretable description of daily operational behavior that can help in selecting appropriate days for a NAS simulation validation.

In this paper, we pursue two coupled interests. First, we design a NAS feature vector that can be used to characterize NAS behavior for comparison purposes. This is analogous to the economic indicators used for performance evaluation of the US economy (e.g., unemployment rate or gross domestic product). Cluster analysis is used to condense an initial collection of 65 aggregate, daily variables into a smaller, more manageable set of variables. Each feature vector then represents NAS performance on its respective day. Second, we use another form of cluster analysis to rank the vectors (and therefore days) by levels of normality.

To understand variable dependencies, we use *cluster analysis* to partition variables into groups so that within each group, the variables display similar behavior. From any of these cluster groups, one may select a single variable as representative. Or, variables within groups may be summed (e.g., the number of Ground Delay Programs (GDPs) and Ground Stops (GSs) summed is GDPs+GSs), thus reducing the number of components to be considered for a NAS feature vector.

The second part of this paper investigates the classification of several "types" of days in the NAS. Once again, we use a cluster analysis approach to pursue this objective. Given the "optimal" NAS feature vector as a basis, we investigate the collection of NAS feature vectors, one vector for each day from Jan. 1, 2000 through Sept. 10, 2001, to identify the natural clusters of types of days. In the analysis, we did not force any partitions. Rather, the data naturally reveals certain types of days in the NAS. In particular, the analysis shows that weather and GDPs play an important role in determining the types of days in the NAS.

The term "cluster analysis" is necessarily broad and encompasses a wide variety of clustering algorithms. Even within a particular statistical endeavor, there can be many ways to cluster the data into meaningful groups. For instance, distance-based cluster algorithms map the variables into *n*-dimensional space, and then check for geometric proximity, using any of a number of metrics. As clusters develop, the trick is how to define distance between multiple objects. Some concept of cluster "center" must be applied.

Cluster analysis is a mature science[1,2,3]. For the most part, clustering algorithms fall into one of three categories:

1. Tree-based clustering, (data are broken into groups, by successive branching on variables),
2. K-means clustering (forms a specified number K of clusters so that there is similar variance within each cluster but dissimilar variance between clusters), and
3. Two-way joining (clusters formed for "cases" and variables at the same time).

In our analysis, we opted for a combination of tree-based clustering and K-means clustering. Our strategy thereby afforded us both robust control over and review of the clustering process.

The best way to understand cluster analysis is through analogy. Suppose someone asks you what your "typical" day is like. Specifically, what time do you get out of bed in the morning? You may reply: "weekday or weekend?" Knowing that your habits differ on these two types of days; it would be meaningless to consider an average across all such days. For if you wake at 7:00 am on weekdays, but sleep in until 10:00 am on weekends, then your average time of rising is 8:30 am. Clearly this statistic is misleading, because there may be virtually no day on which you get up at 8:30 am. Although this fulfills the notion of "typical" in some average sense it lacks a sense of frequency or modality. By contrast, the average rising time of a weekday has meaning in the "typical" sense, since rising times are fairly tightly grouped around 7:00 am on weekdays. Moreover, we have a strong chance of finding a typical weekday (historically speaking) in which you rose very close to 7:00 am. Overall, a major purpose of cluster analysis is to insure that within each cluster, data are actually present close to the mean, and with reasonable frequency.

The purpose of the cluster analysis across NAS feature vectors is to determine whether certain variables should be split into multiple groups with similar behavior. In our rise-and-shine analogy, the cluster analysis would first detect the weekday vs. weekend difference and split out two groups before even asking the question what is typical. Moreover, the analysis might even identify a third category, called "vacation days" with highly erratic rising times. It may seem reasonable to add the vacation days to the weekend category, but the high variance might compromise the integrity of the weekend data set. The purpose of a cluster analysis would be to determine from an objective, scientific standpoint whether or not this is a reasonable thing to do.

Data Associated with Sept. 11, 2001

Data in this study are split distinctly before and after September 11, 2001 (9/11). It is well known that air traffic volumes precipitously dropped immediately after 9/11. Clearly, 9/11 – and a few days thereafter – should be treated as a special event. What about the response and explanatory behavior of the other variables after 9/11? Are the days after 9/11 just low-volume instances of days prior to 9/11? The answers to these questions help us determine whether we should restrict our attention to pre-9/11 data or to span over it.

A Principal Component Analysis (PCA) is a statistical procedure that transforms a number of (possibly) correlated variables into a smaller number of uncorrelated variables called *principal components*. We adopted a form of PCA known as *oblique* PCA and *centroid-based clustering*[4,5] (available within statistical analysis software). With these methods, we found that most of the candidate NAS feature vector variables contribute a similar variation in the pre- and post-9/11 datasets. Nevertheless, we also found some unique differences in the two datasets that require more detailed investigation outside the scope of this effort. So to be cautious, we proceeded only with the pre-9/11 dataset to obtain optimal clusters. For example, the scheduling groups of variables were quite different in the two periods. A complete understanding of the differences between these two datasets remains an open research question outside the scope of the effort recorded in this paper.

Data Sources

As illustrated in **Figure 1**, a total of 65 variables were considered in a cluster analysis. These were taken mainly from the FAA's Operations Network Database (OPSNET), FAA's Aviation System Performance Metrics (ASPM), Bureau of Transportation Statistics (BTS), and Air Traffic Control System Command Center (ATCSCC) quality assurance data sources. The variables captured delay statistics (e.g., en route, terminal), traffic counts, traffic management initiatives (e.g., GDP, GS, miles-in-trail restrictions), and limited weather information (e.g., Instrument Flight Rule (IFR) vs. Visual Flight Rule (VFR) conditions).

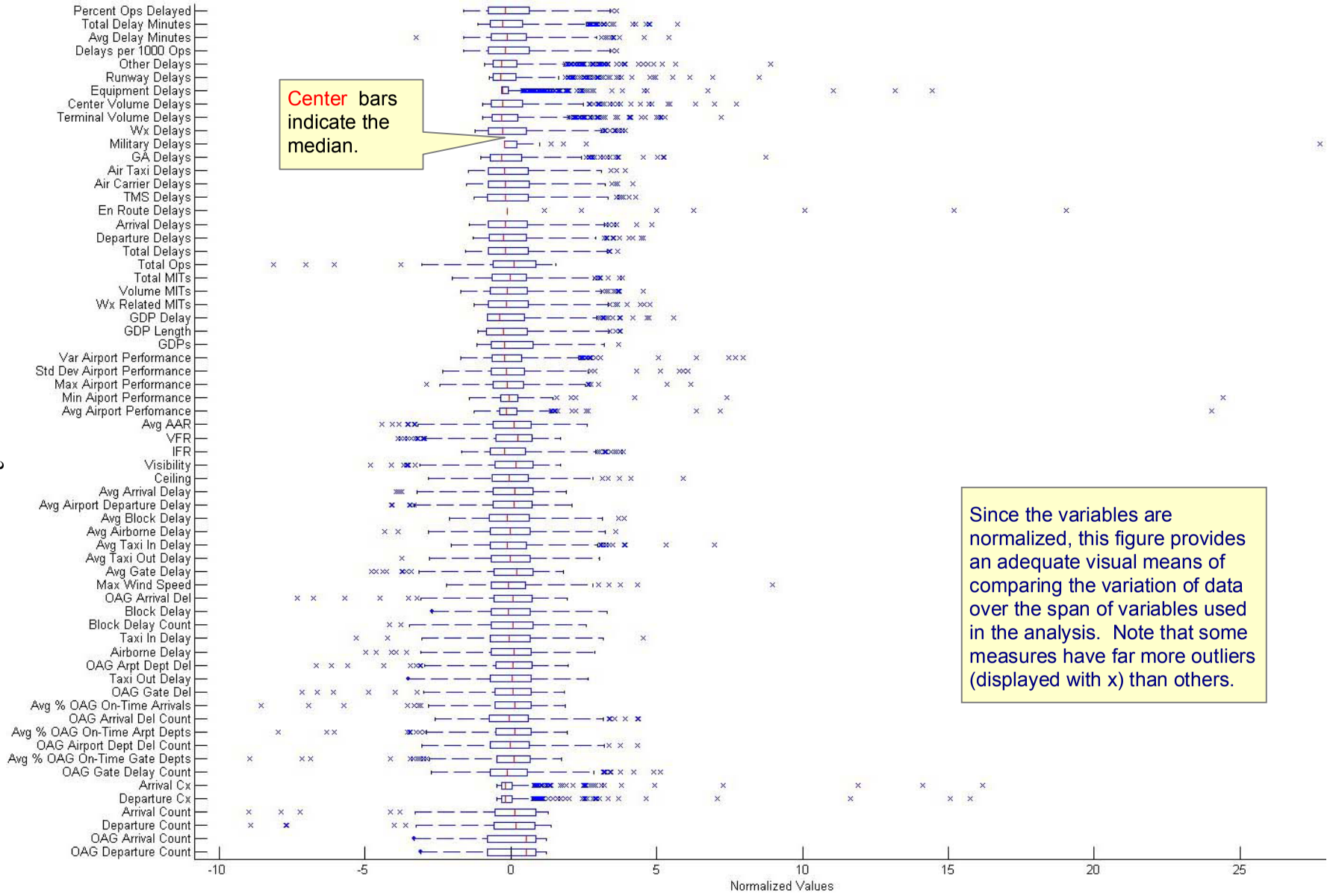

Figure 1.: Analysis of states, control actions, and performance measures of the NAS.

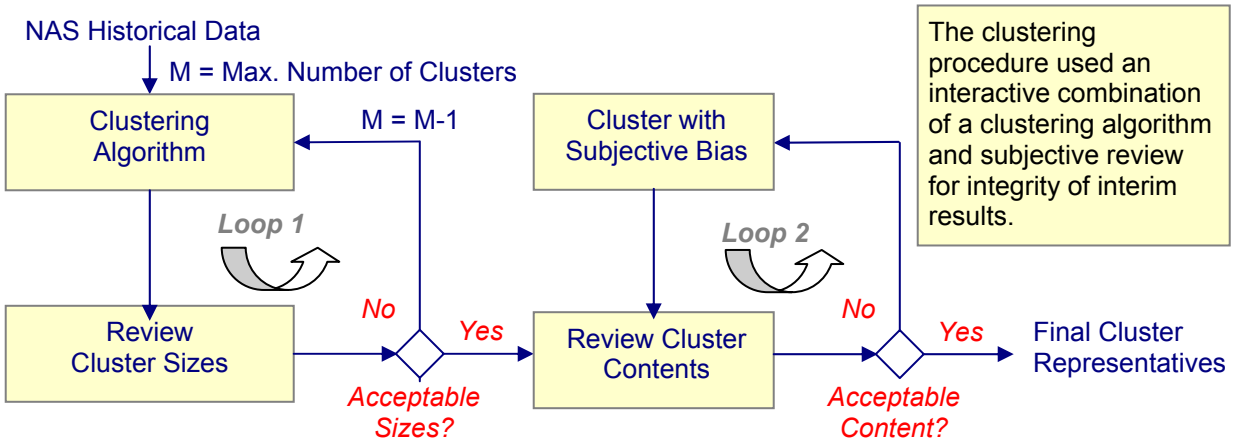

Figure 2: Cluster analysis flow chart (note: Loop 2 was planned for in our analysis but it was not required, that is, the content was determined to be acceptable).

Cluster Analysis Approach

In both the optimal feature vector analysis and the types-of-day analysis, a two-phased cluster analysis was pursued.

The process for identifying the optimal NAS feature vector is depicted in **Figure 2**. The candidate 65 variables were clustered (bundled) by similar statistical behavior to reduce the set of variables to a more intuitive, manageable set. This set of key variables defines what we call the "optimal" NAS feature vector.

The first step in **Figure 2** is an oblique PCA clustering algorithm, which forms an initial clustering. The algorithm was run with an initial setting of the maximum number of clusters. Each cluster was reviewed for uniformity and membership count. In general, a clustering algorithm will return the maximum number of clusters, because its internal objective function can always be reduced by creating one or more singleton clusters. So, the algorithm was run again with the maximum number of clusters decremented by 1. This cycle, Loop 1 in **Figure 2**, was repeated until cluster memberships were deemed reasonable. A good rule of thumb is that each cluster should have at least 2% membership.

Once the cluster membership counts were acceptable, the clusters were reviewed for content. The algorithm clustered variables based on similar statistical behavior (variance). Intuitive knowledge of NAS behavior and variable nuances was applied to ensure that the algorithmic groupings are consistent with known, or suspected, relationships. For instance, one would expect to see daily departure counts and arrival counts to be in the same cluster.

If content of one or more clusters were unacceptable, then the process enters Loop 2. There are two options: to override the clustering algorithm with subjective bias – that is, force certain groupings – or to appeal to another type of clustering algorithm. Given the NAS data of our study, our process never had to enter Loop 2.

Phase I: Clustering to Determine an Optimal NAS Feature Vector. In Phase I, we determined the optimal NAS feature vector. Loop 1 was executed six times. With each iteration, we decided whether certain variables should be eliminated. The criteria for potential elimination of a variable v were:

- v is redundant, i.e., there exists another variable w with an unusually strong correlation with v (hence there is no need for both v and w);
- v is "homeless", that is, it has an extremely weak association with all of the clusters;
- v is essentially constant over time.

In all, we eliminated all but 8 of the 65 variables.

Once the cluster content stabilized and met with our approval (engineering judgment), the most meaningful representative from each cluster was chosen. Since the algorithm outputs an index for each variable – which represents the strength of association with its cluster – the variable with the strongest association was chosen as the representative. Nevertheless, a similar variable may be chosen instead for subjective reasons. We overruled the default selection in only one case.

The clustering process culminated in 8 variable bundles and their representatives, which are listed in **Table 1**.

Table 1: Optimal Variable Cluster Set (further details given in Krozel, Hoffman, et al[6])

Cluster	Cluster Name	Prominent Variable within Cluster	Members
1	Gate Delays	Daily Count of OAG-Based Gate Delays	6
2	Overall Delays	Total Delay Count From OPSNET	14
3	On-time Performance	Daily Total OAG-Based Airport Departure Delay (minutes)	7
4	Traffic Volume	Daily Arrival Count	9
5	Airport Performance Metric	Std Dev of Airport Performance Score (21 ASPM Airports)	3
6	Cancellations	Daily Arrival Cancellations Count	2
7	Volume-related Delays	Total Operation Count From OPSNET	4
8	Weather and GDPs	Total Delay attributed to GDPs (minutes)	11

Gate Delays	*Overall Delays*	*On-Time Performance*	*Traffic Volume*	*Airport Performance Metric*	*Cancellations*	*Volume-Related Delays*	*GDPs*
3490 flights	190 flights	14,500 min.	20,081 flights	5.474	471 flights	47,600 flights	7,480 min.

Figure 3: Feature vector for February 11, 2001.

Each cluster was given a name to convey the major theme of the comprising variables. For each day, the eight corresponding aggregate statistics were compiled, with the intent of performing a Phase II cluster analysis to determine the different "types" of days in the NAS. For instance, the feature vector for February 11, 2001 is shown in **Figure 3**.

Phase II: Clustering to Determine the Types of Days in the NAS. In Phase II, a cluster analysis was performed on the NAS feature vectors passed on from Phase I. The objective in Phase II was to classify the NAS feature vectors for each day Jan. 1, 2000 through Sept. 10, 2001 into groups that naturally described different types of days in the NAS. For instance, if one of the variables were overwhelmingly bimodal, then the algorithm would tend to break the vectors into two groups corresponding to the two (implicit) distributions given by that variable. Without this crucial step, the multi-modal nature of the NAS feature vector components might render the type-of-day classification meaningless.

In theory, a clustering algorithm could break the feature vectors (data points) into any number of clusters. Each cluster would represent a different type of day in the NAS, and within each cluster, we could define typical and atypical days. But, on an intuitive level, it seemed that the number of types of days in the NAS should be relatively small. For instance, a natural decomposition might be six clusters, resulting from three levels of traffic volume, each with two possible levels of weather conditions.

In the Phase II cluster analysis, a centroid-based (K-means) clustering algorithm was used. The overall iterative process was the same as the variable bundling process (**Figure 2**), with two exceptions: (1) the data points are days in the calendar year rather than NAS feature vector variables, and (2) no data points were eliminated.

A relaxed cluster analysis was performed, meaning that we did not interject any subjective biases into the algorithm with a generous upper bound on the maximum number of clusters (20). This resulted in 20 clusters (as expected) but only 10 of these had significant membership – many of the other clusters had only 1 or 2 data points in them. So, we executed the process again with the maximum number of clusters set to 10. This drove the singletons back into the major clusters. This time, only 7 of the 10 resulting clusters had significant membership, so we ran the process one more time with the maximum cluster value set at 7. The resulting 7 clusters had memberships of 62, 183, 104, 68, 16, 182, and 4. Each of these is considered significant (at least 2% of the number of data points). The low membership of 4 days in Cluster 7 was a bit unsettling, but examination revealed that these days were statistical outliers, which are often grouped together in a cluster analysis. This means that there were mainly six major clusters.

Key Variables in Clustering

Satisfied with the resulting cluster membership counts, we proceeded to investigate which, if any, of the variables had been the primary determinant in dividing the data. (If there were no recognizable

pattern, then it would be hard to characterize the clusters as to which types of days they represent.) The X-gobi software tool was used to visually examine the data and clusters from multiple dimensions. In particular, we plotted each of the eight variables against the cluster numbers. A typical scatter plot is shown in **Figure 4**. Each point (m,n) represents a day belonging to Cluster n, which had m hours of GDPs run that day. The data points are of course gathered along their respective cluster lines, but there is no recognizable relationship between cluster number and GDP hours.

In contrast, we found one variable that had an almost perfect relationship with cluster membership: "GDP minutes", which is the number of minutes of ground delay assigned by the FAA during a ground delay program. Each cluster was almost completely characterized by the number of GDP minutes spanned by its members, as seen in **Figure 5**. Aside from a very slight overlap between ranges, the GDP minutes range increases as the cluster number increases. (The notable exception to this association is Cluster 6, which we will discuss shortly.)

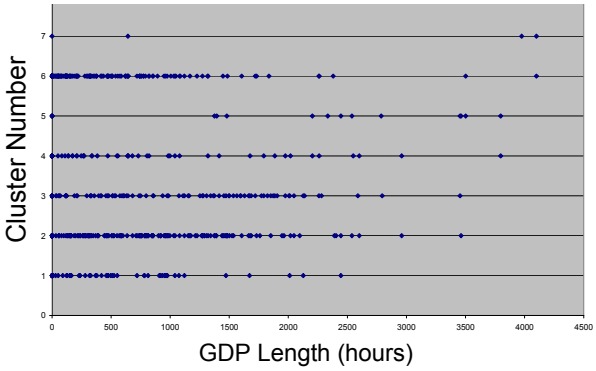

Figure 4: GDP length vs. Cluster.

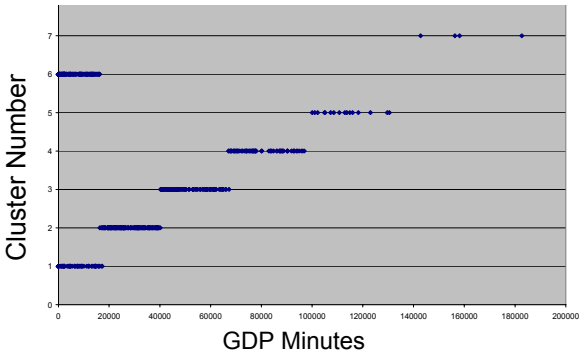

Figure 5: GDP minutes vs. Cluster.

In terms of the types of days in the NAS, we make the following distinction (by GDP minutes):

- Day Type 1: 0 < GDP minutes < 17,302
- Day Type 6: 0 < GDP minutes < 16,236
- Day Type 2: 16,314 < GDP minutes < 40,142
- Day Type 3: 40,257 < GDP minutes < 67,269
- Day Type 4: 67,139 < GDP minutes < 96,931
- Day Type 5: 100,020 < GDP minutes < 130,460
- Day Type 7: 142,770 < GDP minutes < 182,677

Cluster 1 and Cluster 6 clearly share the same GDP minutes range. Cluster 1 overlaps with Cluster 2 by just 988 minutes; Cluster 3 overlaps with Cluster 4 by just 130 minutes; the remaining are non-overlapping.

Statistically, GDP minutes are the single most important variable to consider of the eight representatives when lumping days by similar characteristics. Intuitively, this means that there are generally six types of days in the NAS (seven, if one is willing to count the outliers in Cluster 7), which correspond to the six levels of GDP activity.

Next, we sought to distinguish the two types of low GDP level days. Why did the cluster algorithm choose to break Cluster 1 into two clusters (1 and 6)? Variable *v46*, which is the number of total operations in the NAS for the day, exhibits a bimodal behavior, as illustrated in **Figure 6**.

The algorithm found a more efficient grouping by breaking this cluster into two groups. (A sub-optimal solution would have been to establish 7 levels of GDP minutes, each with a unique range.) By separating Cluster 1 from Cluster 6, the data naturally identified a well-known factor that drives the total number of operations, that is, weekdays (Monday through Friday) tend to have more traffic than weekends (Saturday or Sunday). **Figure 7** further illustrates this point.

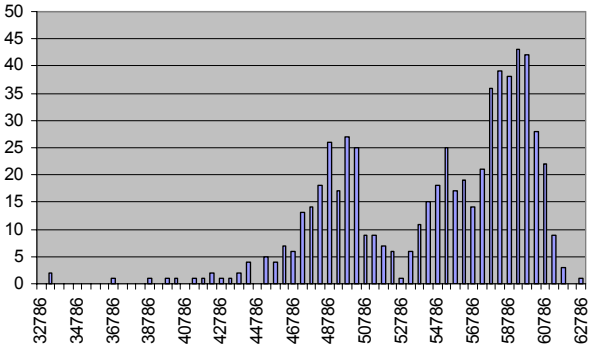

Figure 6: Histogram of total operations count for Jan. 1, 2000 through Sept. 10, 2001. Note the bimodal distribution, attributable primarily to weekday versus weekend traffic.

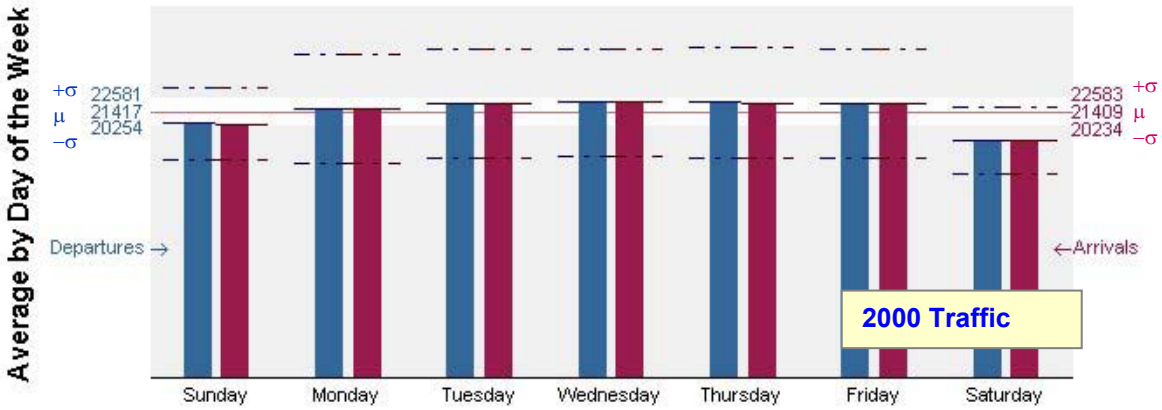

Figure 7. Comparison of traffic based on day of week for 2000.

Traffic volume is a secondary factor in characterizing the type of day in the NAS, next to GDP minutes. This is made clear by **Figure 8**. This is a multi-dimensional projection of the data points. The crosses indicate the days with low levels of GDP minutes (Clusters 1 and 6); the other data points are members of the other five clusters. Note that this group is clearly split into two groups by the bimodal nature of the total operations variable (*x*46).

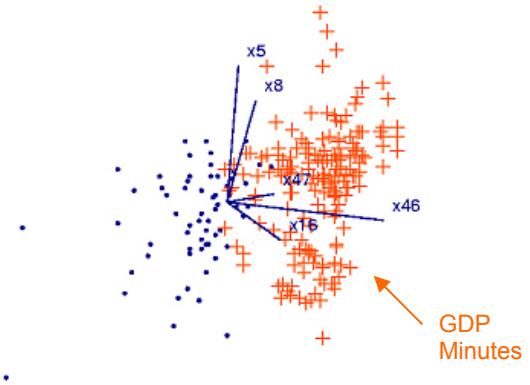

Figure 8: Cluster of low level of GDP minutes.

Having established the six type-of-day clusters, we were able to rank the days within each cluster according to how typical they were, using proximity to the center of the cluster as the criterion. That is, let $\mu=(\mu_1,\mu_2,\ldots,\mu_8)$ be the vector created by setting μ_k equal to the mean of the k^{th} variable of the NAS feature vector, taken over the vectors in a fixed cluster. Mean vector μ is the center of the cluster mass. Then the vector closest to μ was considered to be the most typical day in the cluster. Proximity was defined using a Euclidean-based metric normalized for variance. That is, let $v=(v_1,v_2,\ldots v_8)$ and $w=(w_1,w_2,\ldots w_8)$ be two 8-dimensional vectors. The weighted distance between them is defined as:

$$d(v,w)=\sqrt{\sum_{k=1}^{8}\frac{(v_k-w_k)^2}{\sigma_k^2}}$$

where σ_k^2 is the variance of the k^{th} variable. Without this normalization, proximity would be skewed toward the variables with the larger values.

Table 2 presents the mean vector for each type-of-day cluster and **Table 3** presents the three closest (most typical) days for each type-of-day cluster. The center of the Type 1 type-day cluster is given by row 1 in **Table 2**.

Within each cluster, days can then be ranked according to proximity to the center. The day whose vector is closest to the center of the cluster is considered the most "typical" day in that cluster.

Table 2: Data for the Mean Vector for each Cluster.

Cluster	Count	Variable							
		v8	*v47*	*v16*	*v5*	*v38*	*v7*	*v46*	*v42*
1	62	3737	655	14161	19932	5.722	526.9	46689	6568
2	183	4801	1136	14150	21524	5.455	539.2	54546	28119
3	104	5508	1075	13762	21985	5.229	675.5	54105	51053
4	68	6774	1136	12307	21822	5.145	918.1	54744	79974
5	16	6958	1280	12254	22092	5.183	961.4	56728	112424
6	182	4053	1222	14973	21512	5.801	434.1	56850	5355
7	4	8040	1264	10230	20978	5.339	1008.2	53374	159973

Table 3. Most Typical Day in each Cluster.

Type of Day Cluster	Dates			Distances		
	1st	2nd	3rd	1st	2nd	3rd
1	02-11-01	03-12-00	01-07-01	1443	1799	2643
2	08-03-01	09-05-00	02-17-00	1572	1596	2222
3	10-04-00	06-20-00	06-13-01	2319	2475	3741
4	01-15-01	02-11-00	03-15-01	2930	2935	4103
5	05-22-01	06-16-00	10-27-00	3047	4342	4855
6	07-12-01	05-02-01	03-31-00	1949	2259	2489
7	02-25-01	07-28-00	11-26-00	7035	7780	18252

Interpretation of Results

The results of the clustering algorithm showed that variables *v42* and *v46* were the most critical in separating the day vectors into clusters. We investigate why this is and a range of conclusions.

In scatter plot format, **Figure 9** shows the clustering of data points by variables *v42* and *v46*. This is a projection of the 8-dimensional day cluster data points onto the *v42*-*v46* plane. Each point represents one day; the vertical coordinate is the number of operations for that day, while the horizontal coordinate is the number of GDP minutes.

In **Figure 9,** note that the points on the far left (with low or zero GDP minutes) are divided into an upper group and a lower group, Clusters 6 and 1. This is the effect of the bimodal distribution of *v42*, which we have already seen. The separation is designed to alleviate the debate of which of the two modes (low operations or high operations) is more typical, by breaking them into two clusters, thus making it a moot point.

The objective of the clustering algorithm is to create a fixed number of clusters such that the variance within each cluster is minimized, while the variance between clusters is maximized. Intuitively, this is the same as identifying concentrations of data in multi-dimensional space. Sweeping left to right in **Figure 9**, the breakdown by GDP minutes forms vertical lines of separation. These separations have been made to reduce the variance in the GDP minutes (*v42*) distribution. In the scatter plot, the horizontal separation of the data seems somewhat arbitrary. This is because the frequency of the points with respect to the horizontal axis is obscured.

Consider the frequency distribution of the GDP minutes variable, which is shown in **Figure 10**. The distribution is concave, and heavily skewed to the left. This is the same as saying that in the scatter plot, the number of data points drops off as we move from left to right. The variance of this type of distribution is much greater than that of a classic bell-shaped distribution of equal mass. (In fact, the only way one could rearrange the same mass to have more variance is to evenly divide the mass between the two extreme points.) The intuitive justification for addressing variables with this type of distribution is that they make it the most difficult to answer the question of what is typical. That is, the mean is very far from the mode. (Note that the other variables in the *v42* bundle tended to have similar shapes.)

A natural question is why the cluster analysis process didn't continue making vertical separations of Clusters 2, 3, 4, and 5? If one were to cluster the points solely on the basis of this scatter plot, a strict separation of the points into an upper group and a lower group seems the most natural.

First, as we have pointed out, the greatest payoff is breaking up the horizontal variance, which is not immediately obvious in the scatter plot. A closer examination reveals that the frequency of points drops off as we move from left to right. The magnitude of this drop is not fully appreciated because points on the left are sitting on top of other ones. If one were to draw the points up out of the *v42*-*v46* plane by plotting a third dimension (one of the other variables), then the high variance in *v42* would become clearer. The continuous rotation of data by the X-gobi software made this visually clear.

Second, note that the vertical separation of points in the scatter plot is less pronounced on the right (in Clusters 4 and 5) than it is on the left (Clusters 1 and 6). Overall, the data forms a horseshoe, and the algorithm used an optimization routine to decide where it is no longer profitable to form vertical separations.

The separation of day vectors by variables *v42* and *v46* does not mean that either of these is a more important indicator of the state of the NAS than the other 6 representative variables. It just means that these are the most problematic when trying to determine what a typical day is like. The issue of which of two days is more typical breaks down between clusters, but is preserved within clusters.

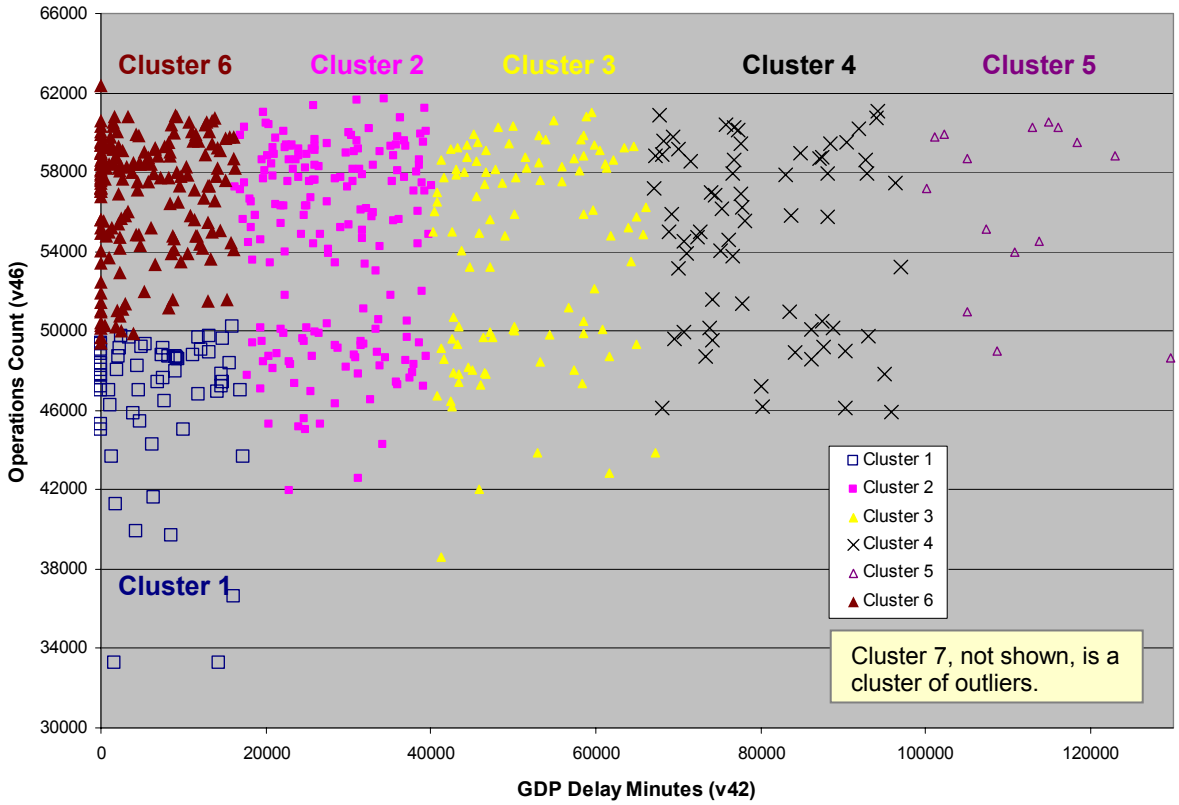

Figure 9: Scatter plot of GDP Delay Minutes vs. Operations Count (Cluster 7 not shown).

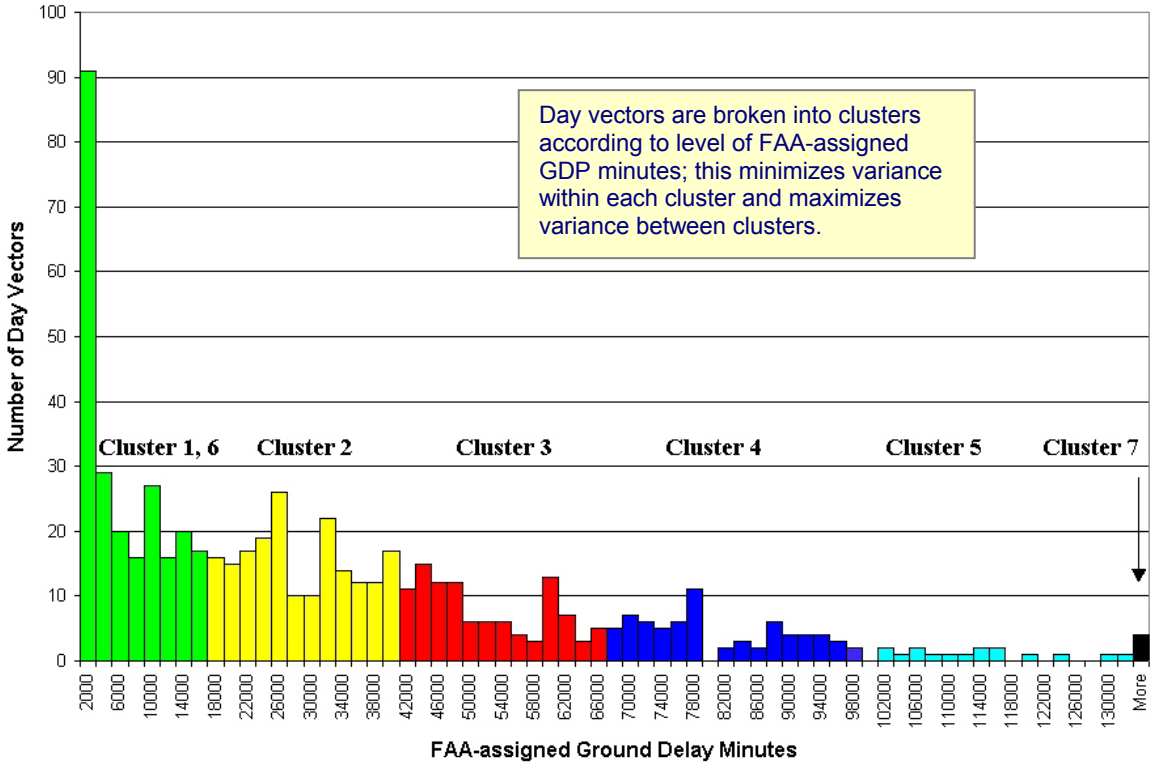

Figure 10: Histogram illustrating the relative cluster locations.

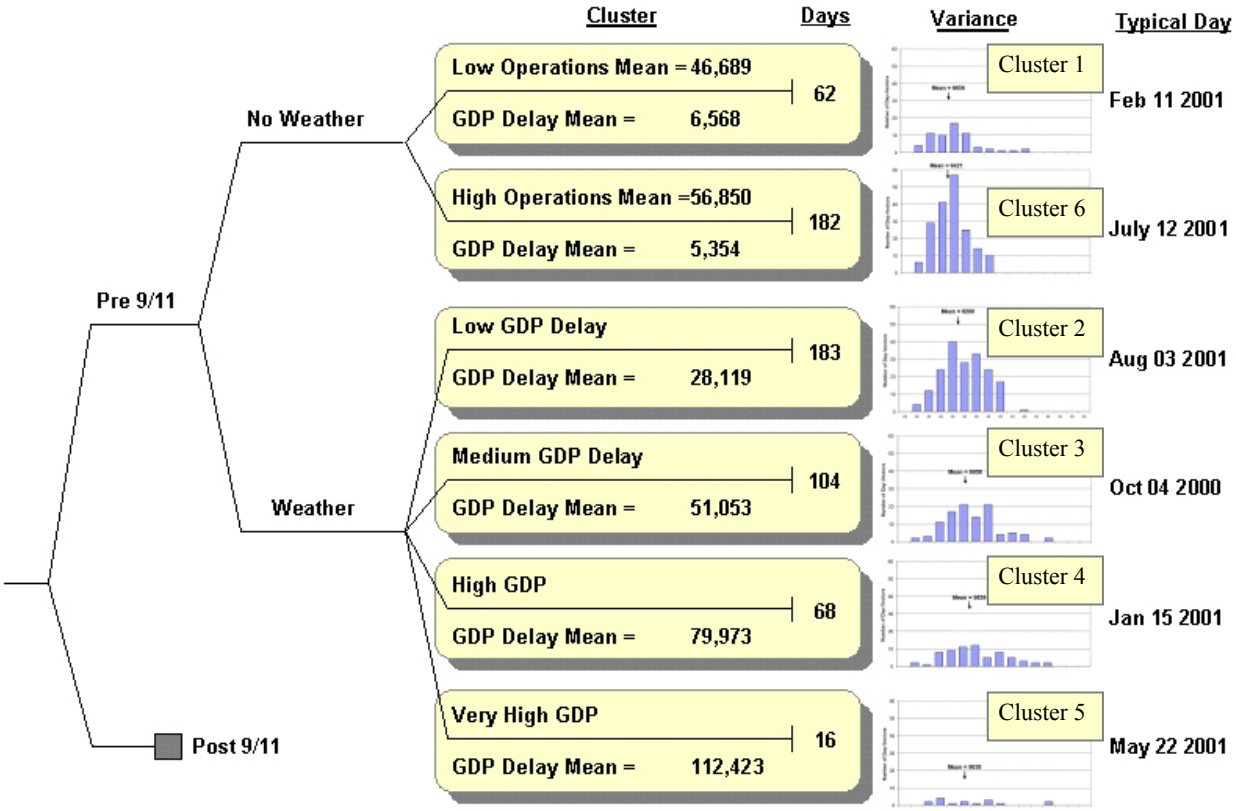

Figure 11: Dendrogram representing the Types of Days in the NAS.

The dendrogram in **Figure 11** summarizes our findings. Reading left to right; a decision at node 1 separates data from pre-9/11 or post-9/11. Since the cluster analysis did not incorporate post-9/11 data, the user is forced into the upper pre-9/11 branch. (The lower branch is included for sake of completeness.) At decision node 2, the user decides whether to choose a data set from the collection of no weather days with very low GDP delay minutes (upper branch) or from the collection of weather days with GDP delay minutes ranging low to high (lower branch). The GDP variable is a surrogate for a collection of ground delay and weather related effects. These branches are labeled "no weather" and "weather" days, which are approximate, descriptive terms. We caution against making comparisons across clusters as to whether one day is more typical than another.

Unfortunately, we were not able to collect any "weather" variables that had much meaning in aviation. For example, variables such as IFR, VFR, and cloud ceilings, do not really help determine what type of day we are having in the NAS. The term "weather" is really a layman's term. What we are really concerned about are meteorological conditions that have an adverse effect on aviation. So, we relied on more indirect indicators of weather, such as GDPs and reduced capacity. In some sense, these are the best weather indicators because they are triggered only if there is weather that adversely affects aviation.

Moreover, the main purpose of this main branch is to avoid making comparisons (with respect to which is more typical) between days with essentially no GDPs and days with some (or many) GDPs. We do not claim that there is literally no weather on the days in the upper branch (though it is a reasonable guide and does not hurt to think of it that way), just that there was no weather that affected the performance of the NAS on those days.

Suppose that the user has selected the "very low GDP" (no weather) branch. At the next branch, we see that there are two types of "very low GDP" days: those with relatively low operations counts (arrivals and departures), and those with relatively high operations counts. A quantifiable definition of these terms is provided in each cluster box by the mean number of operations for that cluster.

Once deciding between these two, there is a unique cluster which houses all days with similar statistical behavior. To the right of the cluster, the most typical day of the cluster has been specified. This would be the optimal data set to consider, meaning that it is most typical. If this day is undesirable for subjective reasons (or if some data elements cannot be collected), then the next most typical day can be

selected. We have ranked the days within each cluster by proximity to the Euclidean center of the cluster; higher indexes indicate a more typical day.

Next, we return to the branch representing the weather impacted days. There are four different types of days to choose from corresponding to days with a low, medium, high, and very high level of GDP minutes. A quantifiable definition of these terms is provided in each cluster box by the mean number of GDP minutes. As described above, the most typical day in the cluster can be chosen as a representative of that cluster (type of day), or another one can be chosen using the cluster ranking.

The overall interpretation of the results is that there are six types of days in the NAS. Each of those has a most typical day (listed in the far right of **Figure 11**). **Table 3** lists additional days with very similar statistical behavior to the most typical days.

Application to NAS Simulation Validations

As for applying these results to the topic of validating NAS simulations, these results indicate that simulation validation sets should consider weather and GDP modeling as a basis for validation data sets. First, if neither weather nor GDP are modeled in the NAS simulation, then the results of our study indicate that there are two types of days that are useful for validating such simulations, which are described by Cluster 1 and 6. If weather and GDPs are include in the NAS simulation – indeed, if weather is included then GDPs must also be included – then, depending on the degree of simulation validation that is desired, there are several choices to be made. One can validate a NAS simulation with modeled weather and GDPs using Clusters 2 through 5 (with special attention to the lower membership size of Cluster 5). Furthermore, trends may be simulated by comparing pairs of clusters, e.g., (2,3) vs. (2,4) vs. (2,5), each having a difference in magnitude of weather and GDP significance in the validation data sets. For a complete validation of a NAS simulation, simulation developers should validate their NAS simulations with at least one validation run from each type of day in the NAS.

Linguistic Descriptions

A more intuitive linguistic description of a cluster can be constructed by examination of its center vector. Each component of the center vector $(\mu_1,\mu_2,\ldots,\mu_7)$ can be mapped into a "high", "medium", or "low" category by considering its distance from the mean of the variable over the entire data set (over all clusters). The categories are established based on the criteria:

Low:	$\mu_k < M_k - \sigma_k$
Medium:	$M_k - \sigma_k \le \mu_k \le M_k + \sigma_k$
High:	$M_k + \sigma_k < \mu_k$

where σ_k is defined as the standard deviation of the k^{th} variable, and M_k represents its mean over all data. This mapping allows for an intuitive description of a given cluster, such as "Low levels of scheduled departures, Medium levels of taxi-in delay", etc.

While these types of linguistic descriptions may help to understand the clusters and the different types of days in the NAS, we caution that the mapping from variable means to high-medium-low categories will probably not be unique. There may be more than one cluster with the same intuitive description. This means that one cannot work backward from the high-medium-low descriptions to create clusters. In particular, one cannot conclude that two days have "similar behavior" just because their high-med-low descriptions are the same. This could corrupt simulation model validation or demonstration efforts.

Conclusions

The first and foremost conclusion of this study is that there is no single day of the year which could be described as a “typical” day in the NAS. One must select the type of day in the NAS first before identifying the most typical day of that type. Hence, we identify a total of six typical days in the NAS, one for each of six representative types of days in the NAS.

We have concluded that a day in the NAS is described by a set of 8 key variables that constitute an optimal NAS feature vector. We reached this point by considering 65 NAS variables in our analysis, which statistically clustered into 8 major bundles, each bundle with a single representative variable. These variables represent the 8 variables that constitute the “optimal” feature vector for the NAS, including: gate delays, overall delays, on-time performance, traffic volume, airport performance metric, cancellations, volume-related delays, and weather/GDP minutes. The number of GDP minutes is the most prominent variable in characterizing the different types of days in the NAS. With the exception of "blue sky days", once the number of GDP minutes is known, a determination of how typical a day is, can readily be made by comparing it to other days with similar GDP minutes. A weighted Euclidean metric (normalized for variance) was used to rank each day within a cluster as most typical to least typical; days closest to this center of the cluster were considered most typical.

Finally, there is an important word of caution learned during the course of this study; namely, data integrity must be considered. Certain data sources were plagued with missing records, typographical errors, incorrectly formatted entries, and poor documentation. This posed a challenge to overcome, as much of the analysis required both well-formatted and complete data sets to be of value. A fair amount of

effort was required to cleanse the data, and this entailed developing software routines that would revise inconsistent records in most circumstances.

Recommendations

Our analysis suggests that validations of low fidelity NAS-wide simulations should mainly focus on the 8 variables of the optimal NAS feature vector. This recommendation will potentially reduce the total quantity of data analyzed in validating a low fidelity NAS simulation. We did not investigate this issue with respect to medium and high fidelity simulations, so we refrain from making a recommendation for validating those types of simulations. When higher fidelity is added to a NAS simulation, more than just the aggregate statistics should be considered for validation. Additionally, one must note that our recommendation assumes that there is no other variable independent of the 8 variables in the optimal NAS feature vector important to a NAS simulation validation. Our recommendation is that NAS simulation validations should consider at least those elements that constitute the optimal NAS feature vector, and if not possible, to attempt to select a substitute from the same cluster set.

The NAS is a very complex system with very many variables that describe it. A very small subset of these variables was studied in our analysis, and of those, the minimal set of variables was determined to define an optimal NAS feature vector. This approach is open to speculation when a new variable that was not in the original set of 65 variables is introduced. While engineering judgment was used to select a set of 8 variables that most likely characterize the NAS behavior, we were limited to variables that are available in historical datasets. Thus, our conclusions are limited to what can be said about how the 65 variables relate to the 8 variables of the optimal NAS feature vector. Caution must be taken when considering new variables outside the set of 65 variables in this study. In such a case, we recommend that a small-scale study be performed to test if the new variable is dependent on one or more of the dominant variables in the optimal NAS feature vector. If the new variable is dependent, then it is not recommended for inclusion in the validation dataset. If the new variable is independent, then engineering judgment should be used to determine if the new variable should be included in a NAS simulation validation.

Future Research

In this study, performance statistics were collected and assessed on a NAS-wide level. An area of future study would be to apply a geographical component to the study. Questions of interest are:

- Can a region of the country serve as an indicator of overall NAS behavior? For instance, if delays are high in the northeast, does this mean that delays are high all over the NAS? Can we collect performance and delay statistics strictly in one region (e.g., the northeast) to assess the overall condition of the NAS?
- In terms of performance metrics, what is the most natural decomposition of the NAS into regions? Does this decomposition coincide with the Air Route Traffic Control Centers (ARTCCs)?
- Are there any local anomalies (e.g., in weather and delays) severe or noteworthy enough to be significant drivers of NAS-wide statistics?
- Our reduction of the size of the NAS feature vector to a set of 8 variables greatly simplified the amount of data that was analyzed in the final cluster analysis. However, we still gather statistics over the ASPM-50 airports whenever possible. This leads us to the following questions: What is the smallest number of airports whose performance is a reasonable surrogate for NAS-wide airport performance? And which airports are these?
- Are there days when a small, local weather disturbance causes big problems? For example, can a small isolated storm over Chicago, IL or New York, NY cause NAS-wide problems? To what degree does fog in San Francisco, CA affect the NAS?

A major theme which was absent – and which could be taken up in future research – is the concept of a cause-and-effect chain. For instance: A, B, and C cause D; B and D cause E; A, B, C, and E cause F. With such a chain, we might identify that certain variables may be treated as dependent variables in one instance, but as independent variables in another.

Acknowledgments

This work supports the NASA Virtual Airspace Modeling and Simulation (VAMS) Project and was performed under Task Order 73 for NASA Ames Research Center's Advanced Air Transportation Technologies (AATT) Project under a subcontract with TITAN Systems Corp., SRC Division.

References

1. Gordon, A., *Classification*, Chapman - Hall, 1999.
2. Hartigan, J., *Clustering Algorithms*, Wiley, New York, 1975.
3. Rand, W., "Objective criteria for the evaluation of clustering methods," *Journal of the American Statistical Association*, Vol. 66, pp. 846-850, 1971.
4. Anderberg, M. R., *Cluster Analysis for Application*. New York: Academic Press, 1973.
5. Harman, H. H, *Modern Factor Analysis*. Chicago University Press, 1976.
6. Krozel, J., Hoffman, B., Penny, S., and Butler, T., *Selection of Datasets for NAS-Wide Simulation Validations*, Technical Report, Metron Aviation, Inc., Herndon, VA, Oct., 2002.